Continuity preserving signal processing

Andringa, Tjeerd

Publication date:
2002

Rijksuniversiteit Groningen

Continuity Preserving Signal Processing

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. D.F.J. Bosscher,
vrijdag 22 februari 2002
om 16.00 uur

door

Tjeerd Catharinus Andringa
geboren op 30 mei 1964
te Leeuwarden

Promotor:	Prof. dr. ir. H. Duifhuis
Beoordelingscommissie:	Prof. dr. L.W.J. Boves Prof. dr. L.P. Kok Prof. dr. L. Schomaker

Continuity Preserving Signal Processing

TABLE OF CONTENTS

Introduction

This thesis is the product of a research project that is simple to formulate: *apply the rich temporal information of a model of the human inner ear (or cochlea) to automatic speech recognition*. This project seemed, and proved to be, an interesting combination of physics, engineering, and cognitive science.

When I started to study speech and auditory processing after my undergraduate study physics, I was struck by the large number of specialisms and subspecialisms. Is this compartmentalization of research the optimal approach for a phenomenon like speech? A definite advantage of a high degree of specialization is that speech and auditory processing are studied in detail, but too much focus on details may prevent the detection of more global regularities. The fact that a speech signal can cause far reaching behavioral responses in less than a second, convinces me that the specialisms ought not only to communicate, but in fact be highly integrated to meet the challenges of the phenomenon of speech.

Unfortunately, research efforts do not reflect this integration, since each specialism seems preoccupied with questions unique for their level of description. Without a solid justification of the compartmentalization of speech science one is forced to be suspicious about the *interpretation* of experiments by one or a few subspecialisms. In particular one might view ideas and theories that are limited to phenomena within a single (or a few) specialism(s) with suspicion because they might represent (subtle) artifacts due to the unnecessary separation of one or more specialisms from the rest of speech processing. It cannot be excluded that these artifacts have prevented the development of a single comprehensive theory of speech processing able to account for the wealth of available experimental evidence.

Another observation that struck me is that, in general, the scientific study of speech processing, as well as engineering approaches of Automatic Speech Recognition (ASR) systems seem to start from a very strong basic assumption: *namely that speech is presented without any background noise.* Yet, as everyone knows, speech is normally produced in uncontrollable situations with multiple sound sources. Consequently source separation ought to be an inevitable ingredient of a speech recognition system. The low noise robustness of modern ASR systems might be hard-coded in these systems by the exclusion of an explicit source separation mechanism and a preference for noiseless signals during design.

A theory of speech, as well as a robust ASR system, might reflect the communicative function of speech as its central premise. *Efficient communication requires that the speech signal changes* slowly *enough for the information carrying features to be detected in a wide range of acoustic environments and* fast *enough to transfer sufficient information.* The balance between these two conflicting demands is likely to determine the main features of the speech signal and is consequently likely to be a suitable starting point for a theory of speech processing.

This thesis is the result of a 10 year period of intermittent research in three groups that each had important influences. In professor Duifhuis' group at the department of *Sensory Biophysics* I was able to work with the model of the human inner ear (or cochlea) that forms the starting point of this work. At the same time I worked as a lecturer for the undergraduate study *applied cognitive science* (*Technische Cognitiewetenschap*, now *Artificial Intelligence*). Here I learned to appreciate the sheer complexity and surprisingly optimal functioning of the human cognitive system. The last years I worked with the research company *Human Quality (HuQ) Speech Technologies*.

Duifhuis' cochlea model has, apart from its careful design and great numerical stability, two special features: it can implement various nonlinearities in a neurophysiologically plausible way and it is continuous in time and place. After a while we decided to focus on the model's *continuity* for a very pragmatic reason: nonlinearities tend to complicate the mathematical description of a problem, while continuity tends to simplify it. After learning more about speech signal processing I realized that the generally applied forms of signal processing where chosen for reasons of mathematical convenience and not to represent the physics of the problem. The symmetry between the continuous nature of the speech production system and the continuity of cochlea seemed too valuable to ignore.

Continuity Preserving Signal Processing (CPSP), possibly in a form similar as developed in this work, is in my opinion the reason why our auditory system is able to keep track of information of multiple concurrent sound sources. Frame-based approaches (FFT's, wavelets, LPC) introduce discontinuities at every transition to a next frame and pose unnecessary restrictions on the trade-off between temporal and frequency resolution. This complicates the analysis of complex signals more than the benefits of a convenient mathematical description warrant. As will be demonstrated in this work, CPSP is not only powerful because it solves problems, but also because it provides alternative signal descriptions in which some of the usual problems of signal processing do not occur. All techniques and representations developed in the course of this work have been patented (Andringa 1999).

Continuity Preserving Signal Processing, as developed in this work, is based on two conjectures. One states that the *auditory system* is optimally useful when it functions *reliably in as many complex acoustic environments as possible*. The other says the same for the speech process. The consequences of these conjectures are discussed in chapter 1. This nontechnical chapter formulates a restrictive framework to guide the development of ideas in later chapters and ends with a formulation of the objectives of this thesis. Chapter 2 starts with an overview of some of the core techniques of CPSP, it continues with the introduction of most of them and demonstrates their potential for ASR applications. Chapters 3 to 5 study special aspects of CPSP. Chapter 6 combines all derived representations to identify coherent areas of the time-frequency plane (termed *auditory elements*) that are likely to represent the most reliable sources of information in the signal. Chapter 7 provides a summary of CPSP and studies its generality. It then returns to the framework presented in chapter 1 by proposing a method to recognize speech sounds in arbitrary acoustic environments. The thesis is concluded by a short analysis of the consistency of CPSP with human performance and an overview of the advantages of CPSP.

Although CPSP is conceptually simple, it deviates in its basis assumptions from modern (speech) signal processing techniques. To allow an optimal presentation, a tutorial form was considered to be more useful than a set of articles. The mathematical prerequisites are minimal, but familiarity with signal analysis is helpful. To help the reader appreciate the differences and special properties of the large number of different representations presented this work, only a single target sentence (namely a Dutch version of *zero one two three*) is used for visual representations.

Although the focus is primarily on speech analysis and speech recognition, this is predominantly because speech is an example of a very complex class of sound. Notwithstanding the single example signal, CPSP is intended as a form of signal processing that can deal with *unknown signals of arbitrary complexity*. HuQ Speech Technologies is founded to take care of the transition of CPSP from a new research tool for the study of complex sounds, to a useful engineering technique that can be readily applied in a wide range of applications. HuQ's products and demonstration systems will prove that CPSP can indeed deal with a large number of qualitatively different sounds. CPSP is in active development and given its huge potential for the analysis of complex sounds it is to be expected that CPSP will develop considerably in the next few years.

Tjeerd Andringa, January 2002